Submitted to the the IEEE International Conference on Bioinformatics and Biomedicine (BIBM09).
draft material -- please do not cite

Towards a Representation of MeSH in RDF

Genaro Hernandez Jr, Ramez Ghazzaoui, Olivier Bodenreider
LHNCBC, National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894
{hernange, ghazzaouir, obodenreider}@mail.nih.gov

Abstract

The Semantic Web provides a framework for the integration of resources on the web, but requires that information sources be represented in a standard form. The Resource Description Framework (RDF) is one such standard for representing resources in the Semantic Web. We describe our approach to represent the Medical Subject Headings (MeSH) in RDF and to validate this representation. We converted MeSH from XML to RDF using XSL transformation and validated our conversion by recreating the original MeSH XML from RDF. We demonstrate that the transformation we performed was lossless. Unlike schemas, the use of unconstrained RDF offers the required flexibility for representing complex terminological structures such as MeSH. Semantic Web mashups will benefit from the availability of MeSH in RDF.

1. Introduction

The Semantic Web provides a framework for the integration of resources on the web, which promises to facilitate information integration and interoperability. One requirement for the Semantic Web is that information sources be represented in a standard form, both syntactically and semantically. Towards this end, the World Wide Web Consortium (W3C) has developed a suite of technologies and standards, including XML, RDF, OWL and SKOS to support the standard representation and processing of information in the Semantic Web.

In the biomedical domain, data integration is an important element of translational research, in which information sources from biological research need to be combined with clinical data sources [1]. To date, however, most biomedical information sources are not available in formats such as RDF and, for this reason, remain difficult to integrate with other data sources in the Semantic Web.

The Health Care and Life Sciences Interest Group (HCLSIG) demonstrated the potential of Semantic Web technologies for data integration and translational research in a 2007 demo [2]. The HCLSIG answered several scientific queries in their demo. One of those queries employed four distinct resources of biomedical information and resulted in a list of potential Alzheimer's disease drug targets.

Central to the HCLSIG demo is that the distinct data sources shared a common representation model for the Semantic Web known as Resource Description Framework (RDF). This alone was not sufficient to enable integration of the data sources, but was an essential step towards integration. The HCLSIG demo proved the feasibility of integrating significant amounts of biomedical data using Semantic Web technologies.

In this paper we describe our approach to represent the Medical Subject Headings (MeSH) in RDF and to validate this representation. This work is a contribution to making existing biomedical resources available on the Semantic Web.

2. Background

2.1. Semantic Web technologies

The Semantic Web is an extension of the current Web and its Semantic Web Stack reveals the technologies that comprise it. These technologies include, but are not limited to, Uniform Resource Identifiers (URIs), the Extensible Markup Language (XML), The Resource Description Framework (RDF) and the SPARQL query language for RDF repositories. URIs are used to identify or name resources on the Semantic Web. XML is a language that enables creation of documents composed of structured data. RDF is the prescribed framework for representing resources in a common format.

RDF describes information in the form of subject-predicate-object triples. This enables information to be represented in the form of a graph. The graph can then

be queried using the SPARQL RDF query language. RDF has several serialization formats including RDF/XML and N-Triples. RDF/XML defines an XML document to encode RDF, while the N-Triple format is a line-based, plain text serialization format for RDF.

2.2. MeSH

The MeSH thesaurus is a controlled vocabulary produced by the National Library of Medicine (NLM) and used for indexing, cataloging and searching for biomedical and health-related information and documents [3]. MeSH consists of three main record types: Descriptor records, Qualifier records and Supplementary Concept records (SCRs). Each record has a unique Identifier. Descriptors, also known as Main Headings, are mostly used to indicate the subject of an indexed item in NLM's MEDLINE and other databases. "Acquired immunodeficiency Syndrome" (D000163) is an example of a Descriptor. Qualifiers, also known as Subheadings, are used for indexing and cataloging in conjunction with Descriptors. An example of a Qualifier is "virology" (Q000821). SCRs are used to index chemicals, drugs, and other concepts for MEDLINE. The nucleoside reverse transcriptase inhibitor Zidovudine, also known as azidothymidine or AZT, is an example of a SCR. A search for Zidovudine in the MeSH Browser [4] retrieves six SCRs. One of these is "zidovudine 5'-monophosphate-mannose-albumin conjugate" (C067831).

2.3. Related Work

Van Assem generated an RDF version of MeSH in Simple Knowledge Organization System (SKOS) RDF Schema [5]. This version of MeSH was used in the HCLSIG 2007 demo and is part of the Bio2RDF mashup system [6] that seeks to help the process of bioinformatics knowledge integration.

SKOS models a thesaurus as a set of SKOS concepts. Instances of the SKOS concept class represent actual thesaurus concepts. Because SKOS is a concept-based model, any feature of a controlled vocabulary that cannot be converted into a concept-based or generic feature is excluded from the representation. Van Assem reports that thesauri can have term-based features, concept-based features or both term-and-concept-based features. Most, but not all, term-based features, in their most basic form, can be converted into concept-based features. Consequently, the conversion of a thesaurus into SKOS RDF can result in the loss of information. MeSH has term-and-concept-based features and some information was lost in the conversion to SKOS. That is, MeSH contains concepts, terms and metadata about concepts and terms. Not all of the information about concepts and terms can be instantiated in the SKOS framework. These differences between the SKOS and unconstrained RDF representations of MeSH are detailed in the Discussion section.

In order to overcome the limitations of SKOS for representing MeSH, we chose the RDF formalism. The contribution of this work is to provide a faithful representation of all features present in the XML version of MeSH in a format directly usable on the Semantic Web. We prove that the transformation process we developed is lossless by performing the inverse transformation from RDF to XML.

3. Materials

MeSH. The 2008 XML version of MeSH [7] and corresponding Document Type Definition (DTD) files were the starting point of our conversion process.

Saxon. The Saxon package is a collection of tools for processing XML documents [8]. Saxon version 9.1.0.2nN was used for our XSL transformations (XSLT). An XSLT is a program written in the declarative language XPath. It specifies how the source XML file is processed and what output is expected.

Closed World Machine (CWM). CWM is a popular Semantic Web program that can perform a variety of tasks [9]. CWM version 1.197 was used with Python version 2.4 to convert N-Triples into RDF/XML format.

ExamXML. ExamXML is a commercial software package for computing and visualizing the difference between two XML files [10]. We used ExamXML version 4.40 in order to compare XML files.

4. Methods

Our conversion process consisted of two main parts: Converting MeSH into RDF and validating the con-version of MeSH into RDF. Figure 1 provides a general overview of the process by which we represented MeSH in RDF and validated our representation.

4.1. Converting MeSH into RDF

The DTD file of each record type was examined in order to understand the structure of the MeSH XML. This informed the creation of an XSLT for each MeSH record. Figure 1 shows that the XSLT was applied to

the MeSH Source XML and resulted in the creation of MeSH RDF in N-Triple format.

4.2. Validating the conversion of MeSH into RDF

In order to validate the conversion of MeSH into RDF, i.e., to prove that the transformation into RDF was lossless, the MeSH N-Triples were used to reconstruct the MeSH XML. The reconstructed MeSH XML was then compared to the source MeSH XML.

4.2.1. Reconstructing the MeSH XML. Figure 1 reveals how the MeSH XML was reconstructed. The MeSH N-Triples were converted to RDF/XML with CWM. To accomplish this step, it was necessary to split the N-Triples into distinct MeSH records because CWM could not handle the large numbers of MeSH N-Triples. The individual MeSH RDF/XML records were modified to include a default namespace since CWM did not include it. This step was necessary so that the XSLT_2 in Figure 1 could be employed. The contents of the individual RDF/XML records were then combined into a single file. An XSLT, specific to each MeSH record type, used the MeSH modified RDF/XML in Figure 1 to reconstruct the MeSH XML.

4.2.2. Comparing the source and reconstructed MeSH XML. Each record from the source MeSH XML file was compared to the corresponding record from the re-constructed XML file using ExamXML. Specific options of ExamXML were selected for the comparison in order to ignore inessential differences such as the order of XML tags.

5. Results

5.1. Converting MeSH into RDF

The 2008 MeSH XML was converted into RDF N-Triples. Figure 2 shows part of a Descriptor record before and after being converted into RDF N-Triples via an XSLT. In Figure 2 we see that the XML representation of MeSH has several tags that include <DescriptorName>. The same information can be represented with RDF N-Triples. For example, an N-Triple in Figure 2 states that the Descriptor record has a “DescriptorName” that is “HIV”. Table 1 shows the number of MeSH records involved in our conversion, the number of N-Triples created and the time required to create the N-Triples.

5.2. Validating the conversion of MeSH into RDF

Comparison of the source and reconstructed MeSH XML indicated both were identical (ignoring inessential differences). This showed that the XSLT used to generate the N-Triples resulted in a lossless transformation of MeSH.

6. Discussion

6.1. Significance

We created a fully automated process to transform the native XML representation of MeSH into RDF, so that it could be used in the Semantic Web. Since MeSH is updated frequently, the transformation process needs to be automated and possibly integrated into the MeSH production environment.

The transformation process we created is also loss-less. By reconstructing the original XML from the RDF representation, we proved that all the information in the source XML has been captured during the transformation into RDF. Applications using the RDF representation of MeSH are thus ensured to access the same information as in the XML version. In other words, the RDF representation is functionally equivalent to its XML counterpart.

6.2. Unconstrained RDF versus SKOS

The conversion of MeSH into SKOS purposely omits a substantial amount of MeSH data and represents neither the complex structure nor complete information of the MeSH thesaurus faithfully. The MeSH structure is far more complex than that of traditional thesaurus and is not amenable to representation with the predefined classes and properties of SKOS. For example, MeSH defines three levels of aggregation of terms: entry term, concept and descriptor, while SKOS only supports terms and concepts. More generally, SKOS imposes a model of a thesaurus and does not allow the flexibility required for representing complex terminological structures such as MeSH.

The aforementioned limitations of a SKOS representation for MeSH are not an issue in our conversion process. By using the N-Triples serialization of RDF, we are able to capture all MeSH in the form of subject-predicate-object triples.

6.3. Unconstrained RDF versus other existing terminology models

Besides SKOS, several models have been developed for biomedical terminologies, often from the perspective of providing specifications for terminology services. This is the case, for example, of the Common Terminology Services (CTS) [11] developed under the auspices of Health Level 7 (HL7). The objective of these models is to provide a unified interface to common features of multiple terminologies.

While useful for the implementation of terminology services in clinical environments, these models do not necessarily capture all features of a given terminology, but rather tend to focus on common, universally useful features. In contrast, our transformation and validation framework guarantees that all the information present in the original XML representation of MeSH is also found in the RDF representation.

Some implementations of CTS allow users to export biomedical terminologies in various formats, including RDF. However, this transformation is not authoritative, as it does not come from the original producer of the terminology. Moreover, the burden of getting the latest version of the terminology lies on the user. In contrast, the version of MeSH in RDF we developed can be made available not only as a downloadable file, but also as a SPARQL endpoint, i.e., a repository accessible over the Internet through the standard query language for RDF. Such endpoints form the basis for linking open data in the data integration framework based on the Semantic Web [12].

6.4. Limitations

While the conversion process is fully automated and does not require human intervention, it may require maintenance. In fact, any changes to the MeSH DTD need to be reflected in the XSLT used to create the RDF. However, such maintenance will be easier than the original creation of the XSLT. Moreover, since the MeSH DTD has been essentially stable in the past years, maintenance of the XSLT is expected to be minimal.

The current version of the RDF representation of MeSH is not final. On the one hand, it needs to be validated by the MeSH development team if it is to become an authoritative version of MeSH. On the other hand, the choice of predicate names and URIs need to be finalized and validated by the Semantic Web community. Base URIs, namespaces and predicate names were chosen somewhat arbitrarily in the development phase, where the focus was on demonstrating feasibility and scalability of the transformation method. However, these elements become important as the RDF version of MeSH is about to be made publicly available. We plan to finalize the representation by next fall.

6.5. Applications

The availability of an authoritative version of MeSH is likely to foster the adoption of this resource in biomedical Semantic Web applications and mashups. Conversely, the lack of such a resource forces Semantic Web developers to create and maintain their own version of the resource.

One motivation for developing an RDF version of biomedical terminologies is that these resources would contribute to seeding a repository of biomedical knowledge, along with knowledge extracted from the biomedical literature by text mining engines and knowledge bases such as Entrez Gene and other resources from the National Center for Biotechnology Information (NCBI). We and others have shown the feasibility and value of creating such repositories by assembling existing resources through Semantic Web technologies [6, 13-15].

The following scenario illustrates the application of the resource we created. Staff members from the Stanford HIV Drug Resistance Database (HIVdb) may want to retrieve PubMed articles that deal with HIV drug resistance but to restrict their search to those articles that mention HIV antiretroviral drugs. This integration of RDF versions of MeSH, PubMed and the VA National Drug File Reference Terminology (NDF-RT) would help refine the search. “Drug Resistance” is a Descriptor with Unique ID D004351. HIV antiretroviral drugs are included in MeSH and the NDFRT. By combining this knowledge one could retrieve the abstracts and PubMed IDs for articles that deal with HIV drug resistance in the presence of antiretroviral drug therapy.

6.6. Lessons learned

The creation of XSLT programs for the transformation of the various types of MeSH records was in fact an iterative process informed not only by knowledge gained from examining the MeSH DTD, but also from the errors observed when attempting to reconstruct the MeSH XML and compare it to the original. The process we developed for creating the lossless transformation of an XML resource into RDF can hopefully provide a model for future

transformations. Overall, the validation part of this project ended up taking a significant portion of the resources and should be factored in when such projects are initiated.

7. Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM) and by the NIH Undergraduate Scholarship Program. Our thanks go to Eric Neumann for encouraging us to create this RDF version of MeSH.

8. References

[1] Ruttenberg, A., T. Clark, W. Bug, et al. "Advancing translational research with the Semantic Web." *BMC Bioinformatics*, Vol. 8 Suppl 3, 2007, pp. S2.

[2] *HCLSIG Demo*. [cited July 10, 2009]. Available from http://www.neurocommons.org/w/images/e/ef/Banff2007Presentation.pdf.

[3] Nelson, S. J., W. D. Johnston, and B. L. Humphreys. "Relationships in Medical Subject Headings (MeSH)." In *Relationships in the organization of knowledge*, pp. 171-84, C.A. Bean and R. Green (eds.), Dordrecht; Boston: Kluwer Academic Publishers, 2001.

[4] *MeSH Browser*. [cited July 10, 2009]. Available from http://www.nlm.nih.gov/mesh/MBrowser.html.

[5] van Assem, M., M. R. Menken, G. Schreiber, et al. "A method for converting thesauri to rdf/owl." In *Proc. Of the 3rd Int'l Semantic Web Conf. (ISWC'04)*, pp. 17-31, S.A. McIlraith, D. Plexousakis and F. van Harmelen (eds.): Springer-Verlag., 2004.

[6] Belleau, F., M. A. Nolin, N. Tourigny, et al. "Bio2RDF: towards a mashup to build bioinformatics knowledge systems." *J Biomed Inform*, Vol. 41, No. 5, 2008, pp. 706-16.

[7] *MeSH*. [cited July 10, 2009]. Available from http://www.nlm.nih.gov/mesh/.

[8] *Saxon*. [cited July 10, 2009]. Available from http://www.saxonica.com/index.html.

[9] *Closed World Machine*. [cited July 10, 2009]. Available from http://infomesh.net/2001/cwm/.

[10] *ExamXML*. [cited July 10, 2009]. Available from http://www.a7soft.com/.

[11] *HL7 Common Terminology Services*. [cited July 10, 2009]. Available from http://informatics.mayo.edu/LexGrid/downloads/CTS/specification/ctsspec/cts.htm.

[12] *Linked Data*. [cited July 10, 2009]. Available from http://linkeddata.org/.

[13] Cheung, K. H., V. Kashyap, J. S. Luciano, et al. "Semantic mashup of biomedical data." *J Biomed Inform*, Vol. 41, No. 5, 2008, pp. 683-6.

[14] Cheung, K. H., K. Y. Yip, J. P. Townsend, et al. "HCLS 2.0/3.0: health care and life sciences data mashup using Web 2.0/3.0." *J Biomed Inform*, Vol. 41, No. 5, 2008, pp. 694-705.

[15] Sahoo, S. S., O. Bodenreider, J. L. Rutter, et al. "An ontology-driven semantic mashup of gene and biological pathway information: application to the domain of nicotine dependence." *J Biomed Inform*, Vol. 41, No. 5, 2008, pp. 752-65.

Record Type	No. Records	No. N-Triples	Time to create N-Triples (min)
Descriptors	24,767	5,542,187	2
Qualifiers	83	9,759	< 1
SCRs part 1	179,704	3,942,609	6

Table 1. The number of MeSH XML records converted to RDF, the number of N-Triples created for each MeSH record type and the time our method required to create the N-Triples of each MesH record type

draft

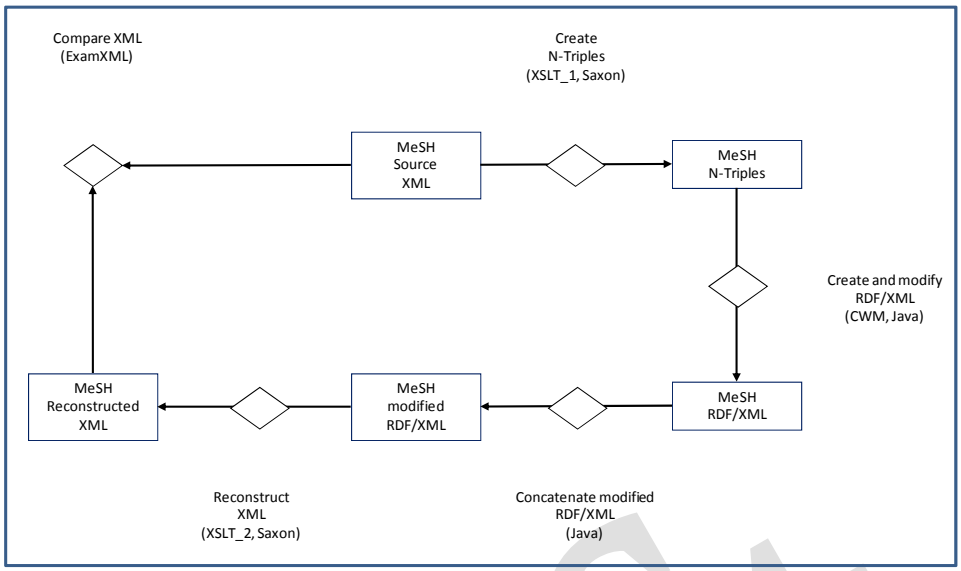

Figure 1. A general overview of the process by which we came to represent MeSH as RDF N-Triples and how we validated this representation by reconstructing the MeSH XML and comparing it to the source MeSH XML.

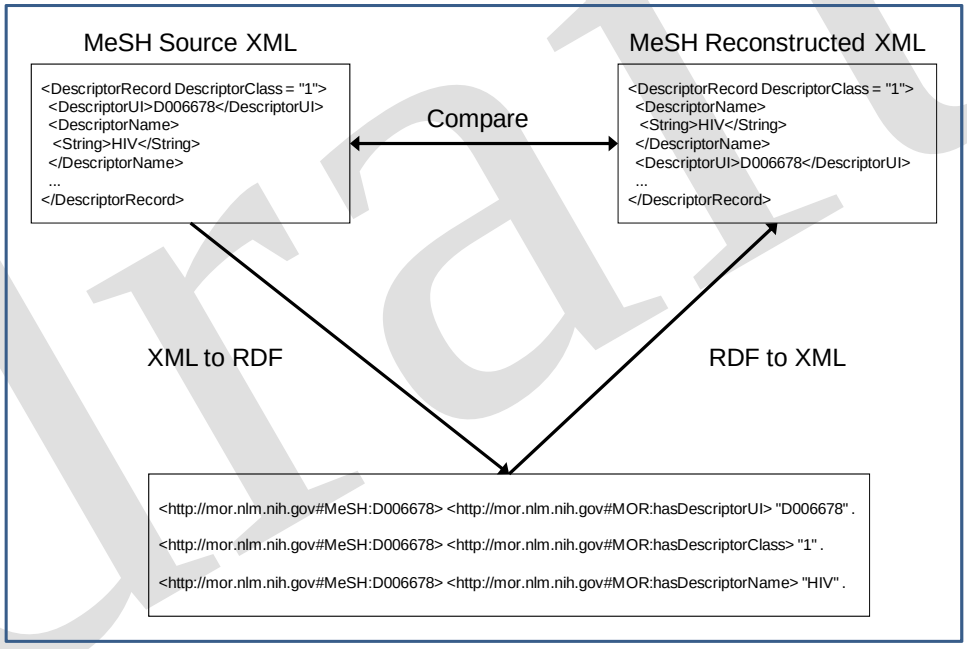

Figure 2. A view of a partial Descriptor record in its original XML format and its RDF N-Triples format that resulted from an XSL transform.

draft